Twitter Event Detection in a City

Martín Steglich(✉), Raúl Speroni, and Juan José Prada

Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay
msteglichc@gmail.com, raulsperoni@gmail.com, prada@fing.edu.uy

Abstract. Large cities and metropolitan areas are complex systems with connections between their environments and individuals. Citizens express themselves daily about events related to the city on the Internet. This information has great value due to its freshness, diversity of points of view and impact on public opinion.

Information technologies allow us to imagine other types of interfaces for communication between people and institutions. Interfaces capable of extracting useful information even if it is not directed to the corresponding institutions.

In this work a framework that combines different techniques for the events extraction in a city from social networks is built. Using the city of Montevideo as a case study and its waste management as a domain, it was possible to correctly identify 94% of the events reported with only 4% false positives.

Keywords: Smart city · Event extraction · Event detection · Social networks · Twitter · Natural language processing · Machine learning

1 Introduction

1.1 Motivation

We are currently facing two social phenomena relevant to the history of humanity: the acceleration of urbanization and the digital revolution. Particularly in Uruguay, access to the Internet from mobile devices and their use to consume and disseminate information by citizens has grown explosively [4,15].

As the population grows, the challenges grow. Increasingly, large cities and metropolitan areas are seen as complex systems with connections between their environments and individuals. Services such as traffic, public transport, waste collection and public safety, among others, require more planning and dynamic decision-making mechanisms that take into account the inclusion of citizen participation processes [3].

Everyday the perspective of the citizens about events related to the city is expressed in blogs and social networks. The information given by the citizens, more and more frequently with the use of mobile devices, does not always go

J. A. Lossio-Ventura et al. (Eds.): SIMBig 2018, CCIS 898, pp. 32–45, 2019.
https://doi.org/10.1007/978-3-030-11680-4_5

through the formal channels provided by the different organizations and nevertheless has great value due to its freshness, diversity of points of view and impact on the public opinion.

The state of the art in disciplines of information technologies such as Natural Language Processing, allow us to imagine other types of interfaces for communication between citizens and institutions. Interfaces capable of extracting useful information for the management of the city even if this information was not generated with the intention of being transmitted formally or was not directed to the corresponding institution. It is possible to imagine that simple complaints or comments on the Internet can become relevant for the management of a city. Therefore it is the object of this work to build event sensing mechanisms able to transform comments on social networks into a valuable resource for the city.

1.2 Objective

The main goal of this work is to develop an information extraction platform from text published on social networks. This solution could be instantiated to obtain events in any city and domain for which quality data exist. The social network Twitter will be used, since it is often the choice of people that wants to make a complaint in real time and because allows the extraction of information easily via API. The solution will seek to identify those tweets in Spanish, which are claims or complaints related to waste management in Montevideo, using techniques of Machine Learning and Natural Language Processing.

The city of Montevideo was chosen for this work because of the quantity and quality of the open data it offers and the waste management of the city is the choice as the domain of the problem. Waste management is one of the most sensitive aspects for the population, so there is a significant amount of complaints in social networks that can serve as a corpus of data for this work.

Fig. 1. Tweets about waste in Montevideo

Although there are centralized mechanisms for the reception of complaints by the Municipality of Montevideo (IM, for its acronym in Spanish), they are not always used by citizens. In many cases, as shown in the Fig. 1, these people prefer the immediacy of Twitter to report a problem in the city.

The solution and techniques proposed in this paper are part of a broader academic work [16].

2 State of the Art

For the purposes of this paper, we can define event as a real-world activity that occurs during a certain period of time in a certain geographic space.

There is abundant previous research on events extraction from written texts. These jobs can be categorized according to the types of events, data sources and methods used [2].

The task of detecting trends in written media generally seeks to identify new issues or topics that have growing importance within the corpus [13]. Following the same line of thought, different techniques of event burst detection in traditional written media has been investigated [6,8–10,12,20].

An example can be seen in [19], Snowsill et al. present an online approach to detect events in news streams based on tests of statistical significance over n-gram frequencies within a time frame. The direct application of these techniques on a large volume of information with noise like that coming from social networks does not seem feasible especially since not all bursts are of interest.

A Twitter-based news processing system called TwitterStand is proposed by Jagan Sankaranarayana et al. in [18]. In that work, a Naive Bayes Classifier is used to determine if a tweet corresponds to a news item or irrelevant information and then a clustering algorithm based on term vectors is used, using Tf-Idf similarity to group the news.

Identifying controversial events that were the origin of public discussions on Twitter is the object of the framework developed by Popescu and Pennacchiotti in [14]. The framework is based on Twitter snapshots. They distinguish between snapshots about events and those irrelevant using supervised decision trees trained on a manually annotated corpus [5].

Another good example of use is seen in [17], where Sakaki et al. used tweets to detect specific types of events such as earthquakes and typhoons. They formulated the event detection problem as a classification problem by training an SVM over manually tagged data to separate positive tweets from the negative ones.

Alqhtani et al. combine in [1] the extraction of Twitter events with the use of Image Mining, a technique that uses Computer Vision and Image Processing concepts so that the images attached to a tweet are considered in the classification.

3 Proposed Solution

It is estimated that an average of 6,000 tweets per second are published in the world [21], the solution must be able to identify those in Spanish, which are claims related to waste management in Montevideo. We will call these tweets "Useful Tweets".

From the moment a tweet is generated by a citizen until it becomes an event to be analyzed by the users of the platform, four stages are covered. The Fig. 2 illustrates how information is enriched in each stage.

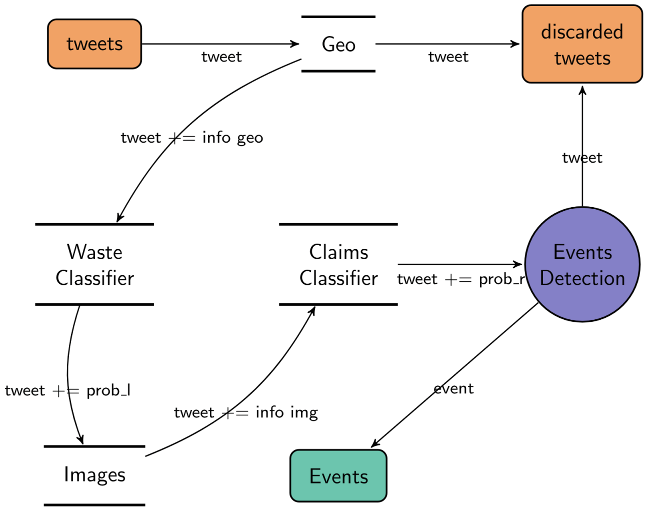

Fig. 2. Module interaction

3.1 Stage 1: Information Retrieval

First of all the platform must recover and store all those tweets that are candidates to be “Useful Tweets”.

Twitter Streaming API is used to collect the tweets. The chosen search criteria was created from the combination of two word lists assembled using observation methods:

- Montevideo related keywords (Montevideo, im, imm, @montevideoim, etc.).
- Waste management related keywords (contenedor, recolector, papelera, basura, etc.).

Each tweet to be recovered, is stored maintaining a common structure with can be used for other data sources that may exist. The stored information keeps the original text, the location, if it exists, and metadata to be used later by the platform.

3.2 Stage 2: Information Enrichment

In this stage, several modules that make up the platform, act. These modules task is to enrich the information obtained in Stage 1. Each module responds to a different strategy to determine if a tweet belongs to the set of “Useful Tweets”. Modules provide new features that enrich the original information and help determine in Stage 3 if a tweet is an event or not.

Georeferencing Module. The aim of the Georeferencing Module is to enrich a tweet during Stage 2 with precise geographic data according to the domain and city chosen.

The input data of the module is the text of the tweet, and some optional geographic data. Twitter provides geographic data if the location of origin of the tweet in enabled, if it is not the module will try to infer location from text. The output is an enriched tweet with a list of possible locations ordered according to an assigned score.

Image Processing Module. The goal of this module is to add information about the attached multimedia content in the tweets, determining if the images attached to a tweet belong to the domain.

For its implementation, the Cloud Vision API service from Google[1] was used, which returns labels that can be used to define whether an image refers to waste management. After some experimentation with labels, the words "waste" and "litter" were chosen to indicate the presence of waste in the image.

Claims Classifier Module. The objective of this module is to determine if a tweet belongs to the set of tweets in Spanish that can be considered claims or complaints.

The training corpus used for the implementation was assembled using a subset of the collection of complaints of the Unique Response System (SUR, for its acronym in Spanish) of the IM. The corpus consisted of a total of 120 thousand claims. These claims are in the catalog of open data of the Uruguayan State and are classified by date and category[2].

Additionally, the Tass General 2015 corpus was used[3], which has approximately 60,000 tweets in Spanish written by different influential personalities from different areas and countries. The themes of the tweets covers politics, football, literature and entertainment [7] and allows the assumption of absence of complaints in its content.

8 thousand tweets that were previously discarded by the georeferencing module were also added because they were originated in other countries. These tweets are of special interest because most of them are not written in Spanish.

The resulting corpus (Claim corpus) was built balancing classes (Claim and Non-Claim) and was divided in train and tests sets for later use.

A classifier was built and an estimator was trained based on the Support Vector Machines (SVM) algorithm using the Scikit-Learn's implementation of LinearSVC[4].

1 https://cloud.google.com/vision.
2 https://catalogodatos.gub.uy/dataset/reclamos-registrados-en-el-sistema-unico-de-reclamos-sur-de-la-intendencia-de-montevideo.
3 http://www.sepln.org/workshops/tass/2015/tass2015.php#corpus.
4 http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html.

In addition, for this classifier the Scikit-Learn's transformers Count Vectorizer (converts the documents into a matrix of tokens count) and TfIdfTransformer (converts a matrix of tokens count into a normalized matrix using the Tf-Idf representation) were used.

The classifier is then a composition of several steps between the transformers and the LinearSVC estimator.

Each transformer as well as the estimators have a set of parameters whose variation can result in a model better adjusted to the particular problem.

To look for the best performance of the classifier it is necessary to train and evaluate the complete model with each combination of parameters.

According to the Scikit-Learn's documentation, it was decided to explore the parameters ngram_range of Count Vectorizer (specifies if the count will be done on unigram, bigrams, trigrams or a combination them) and C of LinearSVC (determine to what extent to favor the search for a hyperplane whose minimum distance to the examples is as large as possible or the search for a hyperplane that separates the examples of different classes as best as possible).

Waste Classifier Module. The objective of this module is to determine during Stage 2 if a tweet belongs to the set of tweets that deal with waste management.

In this module, the SUR corpus was used. By being categorized by domain it is possible to separate the claims that deal with waste from those that do not. Using the domain to which each claim belongs, a corpus (Waste corpus) with approximately 50% of the claims belonging to the Waste class and 50% claims of the Non-Waste class was obtained.

For the implementation of this module a regression model was used. This model is also a composition of the Count Vectorizer and TfIdfTransformer transformers together with an estimator called SVC (another implementation of SVM that allows the prediction of probabilities).

3.3 Stage 3: Information Classification and Events Generation

In Stage 3, the platform must decide whether a tweet is an event or not. Each piece of enriched information incorporated in Stage 2 will be inputs for this decision.

Events Detection Module. The objective of this module is to determine from the results of the Stage 2 modules whether a tweet is an event.

The corpus used for training and testing this module consists of the tweets obtained in the information retrieval stage between September 2016 and January 2017 (Tweet corpus). Each one of those tweets was manually annotated with one of four classes:

- **"VERY_USEFUL"**: Refers to the waste management in Montevideo and contains a specific location.

– **"USEFUL"**: Refers to the waste management in Montevideo and contains some location reference.
– **"BIT_USEFUL"**: Refers to waste management in Montevideo, without location reference.
– **"NO_USEFUL"**: It does not refer to waste management in Montevideo.

For the purposes of this module, a tweet labeled with "VERY_USEFUL" will be considered as an event and others as no events.

For the implementation Random Forests [11] was used, which is a method of assembling Decision Trees. Random Forests starts from the construction of several independent trees during the training and predicts the class of a new example taking the class predicted by the majority of the trees.

The built model is a composition of four transformers, one per module of Stage 2, and a RandomForestClassifier estimator of the Scikit-Learn library. Each transformer takes a tweet and returns a value corresponding to the output of each module. The four values make up the feature vector that will be classified by the model as an event or no event.

3.4 Stage 4: Events Visualization

In the final stage of the system, the events detected in Stage 3 are recovered and displayed to the user. To this end, a webpage was built, showing the map of Montevideo together with the detected events marked according to their location. The location is the best of the ones obtained by the Georeferencing Module of the Stage 2.

4 Results

In this section we present the results obtained by the different modules of the platform.

4.1 Information Retrieval Stage

From the date 27-09-2016 until the date 10-01-2017 the platform retrieved 15,528 tweets, which after being filtered by the georeferencing module, and being categorized and annotated manually, are divided as observed in the Table 1.

4.2 Georeferencing Module

Of the total of imported tweets, the georeferencing module ruled out 7,857 because they are georeferenced tweets located outside the city of Montevideo.

Of the 238 tweets annotated with the label VERY_USEFUL the georeferencing module found correct solutions for 213 of them. Of the remaining 25, 12 had no solution and 13 had wrong solutions.

Table 1. Imported tweets (Tweet corpus)

	Quantity
Discarded by Georeferencing	7,857
Tagged as VERY_USEFUL	238
Tagged as USEFUL	133
Tagged as BIT_USEFUL	2,046
Tagged as NO_USEFUL	5,254
Total	15,528

4.3 Image Processing Module

As explained in Sect. 3.2 for the construction of this module, the Cloud Vision API provided by Google was used. Despite having a small corpus of information pieces with attached images, the results obtained by this module shown in the Table 2 are favorable.

Table 2. Results about tweets with images

	Precision	Recall	F1-score	Examples	% of the corpus
Without-Garbage	0.86	1.00	0.93	840	66.5%
With-Garbage	1.00	0.68	0.81	423	33.5%
Total	0.91	0.89	0.89	1263	100%

To obtain the results, the collected tweets that had an attached image were taken into account. These tweets were inspected manually to determine how many were incorrectly classified by the module. The resulting confusion matrix is shown in Table 3.

Table 3. Confusion matrix over tweets with images

		Estimated Class		
		Without-Garbage	With-Garbage	
Real Class	Without-Garbage	840	0	*False Positives*
	With-Garbage	135	288	
		False Negatives		

In the final results, and as can be seen in the models generated in the Subsect. 4.6 to classify a tweet as an event, it appears that the influence of this module in the definitive classification is low; In general, the images illustrate a reality expressed in the text, causing some redundancy between modules. This redundancy is most noticeable when configuring tweet extraction to work with keywords related to the domain. In a different situation, for example making an extraction of all the tweets in Spanish or all those related to Montevideo, the Image Processing Module would take more relevance in the final result.

4.4 Claims Classifier Module

The evaluation of this module on the test set (33% of the Claim corpus, 42,664 examples) gives a precision and a recall of 1.0, for both categories (Claim and Non-Claim).

The results on the test set are exceptionally good, nevertheless it is good to remember that the model will be used on tweets and not on claims of the SUR system. In Table 4 the indicators can be seen for the Tweet corpus and in Table 5 the confusion matrix can be found. In this case, although the recall is an acceptable 84%, the accuracy is only 14% due to the large number of false positives.

Table 4. Results on the Tweet corpus

	Precision	Recall	F1-score	Examples	% of the corpus
Non-claim	0.99	0.83	0.91	7433	97.53%
Claim	0.14	0.84	0.24	238	3.12%
Total	0.97	0.83	0.88	7621	100%

Table 5. Confusion matrix for the Tweet corpus

		Estimated Class		
		Non-Claim	Claim	
Real Class	Non-Claim	6183	1250	*False Positives*
	Claim	38	200	
		False Negatives		

4.5 Waste Classifier Module

For the implementation of this module a regression model that assigns probabilities of belonging to the Waste class for each tweet was chosen. This type of models allows to vary the threshold above which a document will be considered within a given class.

As an example, taking the threshold = 0.5 and evaluating this module on the test set (33% of the total corpus, 42,076 examples) we obtain an accuracy of 0.97 and a recall of 0.98 for both categories (Waste and Non-Waste).

As in the Subsect. 4.4 the results on the test set of the Waste corpus are very good, but it is important to consider that the model will be used to classify tweets and not claims originated in the SUR system. It can be seen in the Table 6 the indicators when evaluating the model on the Tweet corpus and in the Table 7 the confusion matrix for the same set can be found. In a similar way to the previous section a good recall is obtained for the class sought but a very poor precision, this is caused by the high number of false positives.

Table 6. Results on the Tweet corpus, threshold = 0.5

	Precision	Recall	F1-score	Examples	% of the corpus
Non-waste	0.99	0.42	0.59	7433	97.53%
Waste	0.05	0.90	0.09	233	3.12%
Total	0.96	0.44	0.58	7671	100%

Table 7. Confusion matrix for the Tweet corpus, threshold = 0.5

		Estimated Class		
		Non-Waste	Waste	
Real Class	Non-Waste	3128	4305	*False Positives*
	Waste	23	215	
		False Negatives		

In the Subsect. 4.6, the importance of both domain classifiers in the final classification of a tweet as an event is measured. Even with good individual recall the classifiers seem to have little participation in the final classification, especially the waste classifier. This behavior is due in part to the fact that tweets are retrieved by keywords related to waste, a scenario in which the module has little influence on the final result.

4.6 Events Detection Results

The parametric adjustment was made comparing several alternatives favoring the recall by means of a cross validation evaluation with five partitions.

The outputs of the Stage 2 modules acted as attributes for the Random-Forests model used. Scikit-Learn allows to measure the importance of each feature in the resulting model. The attributes of the model trees were evaluated and the values are seen in the Table 8.

Results for the test set of the Tweet corpus can be seen in the Table 9 and the results and the associated confusion matrix for the complete Tweet corpus are found in the Table 10 and the Table 11. According to this data, the probability

Table 8. Parameters

Feature	Importance
Georeferencing module	0.740
Claims classifier module	0.175
Classifier waste module	0.065
Image processing module	0.018

Table 9. Events detection results for the test set of the Tweet corpus

	Precision	Recall	F1-score	Examples	% of the corpus
No event	1.00	0.96	0.98	2459	97%
Event	0.40	0.95	0.56	73	3%
Total	0.98	0.96	0.97	2532	100%

Table 10. Events detection results for the complete Tweet corpus

	Precision	Recall	F1-score	Examples	% of the corpus
No event	1.00	0.96	0.98	7433	96.9 %
Event	0.44	0.94	0.59	238	3.1 %
Total	0.98	0.96	0.97	7671	100 %

Table 11. Confusion matrix for the complete Tweet corpus

		Estimated Class		
		No Event	Event	
Real Class	No Event	7144	289	*False Positives*
	Event	15	223	
		False Negatives		

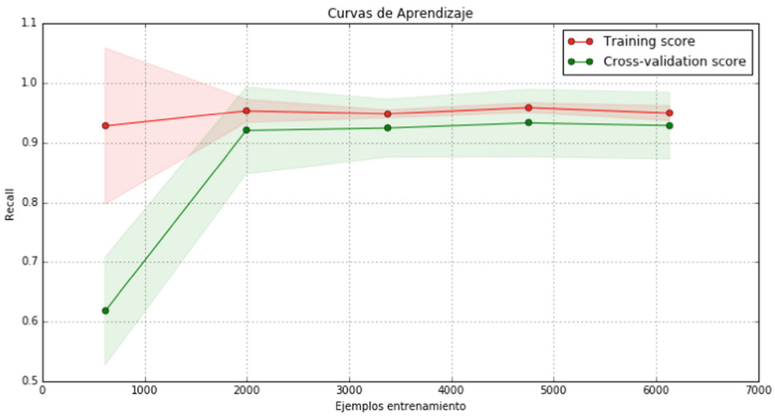

Fig. 3. Learning curves

of classifying an event correctly is 94% (recall) and the probability of incorrectly classifying an event, that is, obtaining a false positive is 4% (fall out).

It is clear that the number of positive instances for the "event" label is low and therefore the results should be taken with caution. However, as seen in Fig. 3 the recall difference between the validation and training corpus starting from certain number of examples seems stabilize, which indicates the ability of the model to generalize predictions.

It is interesting to note that both the classifiers of Stage 2 and the event classifier always sought to favor the recall over the precision in the training stage. The decision follows from the concept that it is better to have fewer false negatives regardless of the number of false positives, that is: from the perspective of city management it matters more to find as many events as possible, even if this means finding more instances that are not events.

5 Conclusions

Everyday more and more citizens interact with each other and with different organizations through technology and social media. In this new era of information exchange, it can be observed that the perception that can be obtained about what happens in a city is more diverse, integral and exact than it used to be.

Artificial Intelligence has taken a great leap in terms of its practical applications thanks to the abundant amount of data generated by Internet users. New tools and algorithms that are applicable to areas so far unrelated to these technologies are released on a daily basis. Industry and public services are examples where diverse techniques of Artificial Intelligence such as Natural Language Processing and different methods of automatic learning can have a great impact.

The generation of data is the consequence of the massive use of the Internet and the main condition for the development of AI applications that are really useful. The availability of quality open data on different aspects of a city is what makes possible research like the one presented in this work. The fact that public institutions have good policies about open data, as in the case of IM in Uruguay, is what allows to move forward in the investigation and analysis of this data for the improvement of processes and services that could benefit citizens.

The platform built was fed with 15,528 tweets during a period of 105 days, these tweets were retrieved by keywords associated with the city of Montevideo and the chosen domain: Waste management. Of the 234 tweets manually marked as possible events, the platform identified 94% correctly while getting only 4% false positives (289 tweets).

The results obtained show that it is possible to identify city events associated to the domain for most of the tweets recovered that complied with the conditions. For these events, it is possible to have a precise geographic location associated to elements of the urban furniture that are the object of the complaint or claim.

The identification of events, located in time and space, from the diverse perspective of citizens and with aggregate information such as photographs or videos allow a level of analysis that would make possible, for example, the detection of problematic geographic areas, the identification of patterns, or sentiment analysis for any given public service.

Going one step further, if each event identified by the platform could be assigned directly to a team for its resolution, and that team could report the result through images directly to the platform, it would close the communication circle with the citizen.

It is in this context of technological advances, and after analyzing the results of this work that we can conclude that it is possible to build sensing mechanisms of social networks that can become a new type of interface between citizens and public organizations. These interfaces would allow a greater participation of citizens in a familiar terrain: social networks, and could offer a more transparent resolution of some claims. These interfaces could become a relevant aspect of the transformation of a city into a smart city.

References

1. Alqhtani, S.M., Luo, S., Regan, B.: Fusing text and image for event detection in twitter. arXiv preprint arXiv:1503.03920 (2015)
2. Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. Comput. Intell. **31**(1), 132–164 (2015)
3. Bouskela, M.: La ruta hacia las smart cities, January 2017. http://goo.gl/3Y0tWK. Accessed 03 June 2018
4. ICT Facts: Figures-the world in 2015. The International Telecommunication Union (ITU), Geneva (2015)
5. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)
6. Fung, G.P.C., Yu, J.X., Yu, P.S., Lu, H.: Parameter free bursty events detection in text streams. In: Proceedings of the 31st International Conference on Very Large Data Bases, pp. 181–192. VLDB Endowment (2005)
7. García Cumbreras, M.Á., Martínez Cámara, E., Villena Román, J., García Morera, J.: Tass 2015-the evolution of the spanish opinion mining systems (2016)
8. Goorha, S., Ungar, L.: Discovery of significant emerging trends. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 57–64. ACM (2010)
9. He, Q., Chang, K., Lim, E.P.: Analyzing feature trajectories for event detection. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 207–214. ACM (2007)
10. He, Q., Chang, K., Lim, E.P., Zhang, J.: Bursty feature representation for clustering text streams. In: SDM, pp. 491–496. SIAM (2007)
11. Ho, T.K.: Random decision forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282. IEEE (1995)
12. Kleinberg, J.: Bursty and hierarchical structure in streams. Data Min. Knowl. Discov. **7**(4), 373–397 (2003)
13. Kontostathis, A., Galitsky, L.M., Pottenger, W.M., Roy, S., Phelps, D.J.: A survey of emerging trend detection in textual data mining. In: Berry, M.W. (ed.) Survey of text mining, pp. 185–224. Springer, New York (2004). https://doi.org/10.1007/978-1-4757-4305-0_9

14. Popescu, A.M., Pennacchiotti, M.: Detecting controversial events from twitter. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1873–1876. ACM (2010)
15. Radar, G.: Perfil internauta uruguayo 2016 - resumen ejecutivo, January 2017. http://www.gruporadar.com.uy/01/wp-content/uploads/2016/11/El-Perfil-del-Internauta-Uruguayo-2016-Resumen-Ejecutivo.pdf. Accessed 03 June 2018
16. Raúl Speroni, M.S.: Extracción de eventos en una ciudad a partir de redes sociales. https://www.fing.edu.uy/inco/grupos/pln/prygrado/Informe_Speroni_Steglich.pdf. Accessed 19 July 2018
17. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM (2010)
18. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: news in tweets. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 42–51. ACM (2009)
19. Snowsill, T., Nicart, F., Stefani, M., De Bie, T., Cristianini, N.: Finding surprising patterns in textual data streams. In: 2010 2nd International Workshop on Cognitive Information Processing, pp. 405–410. IEEE (2010)
20. Wang, X., Zhai, C., Hu, X., Sproat, R.: Mining correlated bursty topic patterns from coordinated text streams. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 784–793. ACM (2007)
21. www.internetlivestats.com/: Twitter usage statistics. http://www.internetlivestats.com/twitter-statistics. Accessed 1 June 2018